Self-critiquing models for assisting human evaluators

Year: 2022 | Citations: 352 | Authors: W. Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Ouyang Long

---

Abstract

We fine-tune large language models to write natural language critiques (natural language critical comments) using behavioral cloning. On a topic-based summarization task, critiques written by our models help humans find flaws in summaries that they would have otherwise missed. Our models help find naturally occurring flaws in both model and human written summaries, and intentional flaws in summaries written by humans to be deliberately misleading. We study scaling properties of critiquing with both topic-based summarization and synthetic tasks. Larger models write more helpful critiques, and on most tasks, are better at self-critiquing, despite having harder-to-critique outputs. Larger models can also integrate their own selfcritiques as feedback, refining their own summaries into better ones. Finally, we motivate and introduce a framework for comparing critiquing ability to generation and discrimination ability. Our measurements suggest that even large models may still have relevant knowledge they cannot or do not articulate as critiques. These results are a proof of concept for using AI-assisted human feedback to scale the supervision of machine learning systems to tasks that are difficult for humans to evaluate directly. We release our training datasets, as well as samples from our critique assistance experiments.